Open Access Library Journal
2025, Volume 12, e14353
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Research on Machine Translation Quality Evaluation of Academic Paper Abstracts: A Case Study of Foreign Language and Literature Papers

Shiyu Jia

School of Foreign Languages, East China University of Science and Technology, Shanghai, China
Email: 1014626137@qq.com

How to cite this paper: Jia, S.Y. (2025) Research on Machine Translation Quality Evaluation of Academic Paper Abstracts: A Case Study of Foreign Language and Literature Papers. *Open Access Library Journal*, **12**: e14353.
https://doi.org/10.4236/oalib.1114353

Received: September 26, 2025
Accepted: November 16, 2025
Published: November 19, 2025

Copyright © 2025 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Machine translation as one of the important applications in the field of artificial intelligence, plays a crucial role in cross-lingual communication and information transmission. However, the quality of machine translation systems directly affects the accurate conveyance and understanding of information. This research aims to investigate the translation quality of different machine translation systems in translating abstracts of foreign language and literature papers. In this study, we selected Youdao Translate, DeepL, ERNIE Bot, and ChatGPT-3.5 as the research objects for machine translation, and evaluated and analyzed the academic paper abstracts translated by these four machine translation tools, exploring their performance in translating academic paper abstracts.

Subject Areas

Linguistics

Keywords

Machine Translation, Translation Quality, Paper Abstract, Evaluation

1. 引言

计算机技术在近些年来快速发展，其应用已渗透到了人们日常生活的方方面面。机器翻译作为计算机技术的热门应用之一，在跨语言交流与信息传递中发挥着日益关键的作用。近年来，随着神经机器翻译(Neural Machine Translation, NMT)与大语言模型(Large Language Models, LLMs)的快速发展，机器翻译的质量与适用场景得到了显著拓展。然而，不同机器翻译系统在处

理专业性较强、逻辑结构严谨的学术文本时仍存在差异，其质量直接影响信息的准确传达与学术成果的国际传播。学术论文作为学术成果的集中体现，是研究人员进行知识探讨与创新的核心载体。其中，摘要作为论文内容的凝练，在研究成果的传播与理解方面起着至关重要的作用。然而，目前国内许多学生在进行摘要翻译时，往往缺乏足够的重视，直接采用机器翻译进行转换，导致译文出现诸多不规范现象，影响了学术交流的有效性。

本研究以外国语言文学领域 CSSCI 期刊论文的中英文摘要作为研究语料，评估不同机器翻译系统在翻译外国语言文学类论文摘要时的表现，以期丰富相关领域的研究，并为学术翻译实践提供参考。

2. 文献综述

2.1. 机器翻译

机器翻译(Machine Translation, MT)也被称为自动翻译，指利用机器把一种自然语言文字翻译成另一种自然语言文字，具有耗时短、成本低等特征[1]。随着个人计算机的普及以及智能手机的广泛普及，在线机器翻译平台因其操作简便、响应迅速且具备一定准确性，已成为大众跨语言沟通的重要工具。近年来，机器翻译技术经历了从基于规则的翻译、统计机器翻译到神经机器翻译的演进，尤其是大语言模型如 ChatGPT、文心一言等的出现，进一步推动了生成式翻译能力的发展，使其在语义理解与上下文连贯性方面展现出新的潜力。然而，机器翻译的准确性尚未完善，国内外学者围绕其展开了大量研究。

国外对机器翻译的研究主要集中在计算机科学、信息工程、材料科学以及人文社科等领域。Hovy 在 2002 年提出了一个机器翻译评估框架(a Framework for Machine Translation Evaluation)，它将用于评估机器翻译的质量模型与系统的目的和上下文联系起来[2]。Haque 等人评估了机器翻译系统在术语翻译方面的表现，揭示了基于短语的统计机器翻译和神经机器翻译在术语翻译领域的优势、劣势和相似之处[3]。Hudelson 和 Chappuis 二人探讨了潜在的对语音翻译在临床医学应用中有利或不利的因素[4]。Koplenig 和 Wolfer 二人基于涵盖 1293 种语言的语料库数据训练计算机模型，发现使用人数越多的语言对于计算机语言模型而言往往更难学习，揭示了语言学习难度与说话者人口规模之间的关联[5]。

国内早期研究多集中于机器译文与人工译文的质量对比，黎斌和唐跃勤以英语特殊句型 There be 句型为切入点，以五种英汉全文机器翻译软件作为测试对象，着重探讨英汉机器翻译中特殊句型的对应与不对应的调整问题[6]；黄海英和冯剑军对五个不同的机器翻译软件进行了评估打分[7]；潘幼博在 1990 年提出机器翻译系统还远远不够完善，只适用于科技资料的翻译[8]。2000 年后，一些学者陆续开始探讨提升机器翻译质量的方式与途径，魏长宏和张春柏提出译后编辑作为机器翻译系统的有机组成部分，有助于提高译文质量和人工译校效率[9]。崔启亮和李闻基于科技文本的英汉机器翻译，确定了译后编辑中的错误类型[10]。冯全功和刘明尝试在已有相关研究的基础上，

构建包含认知维度、知识维度和技能维度的译后编辑能力三维模型，并阐述各个维度的具体构成与表征[11]。冯志伟与张灯柯则强调机器翻译研究需融合语言学知识与常识，主张把基于语言大数据的连接主义方法和基于语言规则与常识的符号主义方法相结合，指出机器翻译将成为人工翻译的好朋友和得力助手，机器翻译和人工翻译应当和谐共生，相得益彰[12]。

2.2. 热门机器翻译工具

有道翻译是网易公司开发的一款翻译软件，其最大特色在于翻译引擎基于搜索引擎与网络释义，能够动态获取并更新词条解释。该软件自推出以来不断进行更新与完善，目前最新已推出有道翻译官 4.1.12 版。

DeepL 翻译器由一家德国公司于 2017 年推出，是一款基于神经网络的机器翻译工具，目标是通过使用人工智能技术消除语言障碍。

百度翻译在 2011 年上线基于互联网大数据的机器翻译系统，后又在 2015 年上线全球首个互联网神经网络翻译系统。2019 年，百度发布大语言模型文心大模型 ERNIE 1.0，经过不断优化，现已更新至文心一言 V2.5.3 版本。

ChatGPT (全名：Chat Generative Pre-trained Transformer)是 OpenAI 研发的人工智能技术驱动的自然语言处理工具，拥有语言理解和文本生成能力，能够进行翻译任务。

2.3. 常用机器翻译质量自动评估指标

BLEU (Bilingual Evacuation Understudy)是国际上机器自动翻译评价系统的流行指标，以翻译文本的 N 元组(N-gram)也就是译文中连续的词的个数为出发点，计算被测译文与参照译文间相似度及距离，数值范围 0~1，数值越高，翻译质量越好；翻译编辑率 TER (Translation Edit Rate)通过统计机器译文修改为参考译文的后编辑次数，来分析机器译文的质量，机器译文所需的后编辑次数越少代表译文质量就越高，数值范围 1~0，数值越低，翻译质量越好；METEOR (Metric for Evaluation of Translation with Explicit Ordering)指标评估时主要以词语为单位，考虑模型生成语句中与正确参考语句匹配的不考虑位置的词语，将词根、同义词、功能词、单词顺序等情况纳入评估范围内，数值范围 0~1，数值越高，翻译质量越好[13]。

3. 研究设计

3.1. 语料选择

学术论文逻辑性强，结构严谨，层次分明，文字简练。同时，期刊上发表的学术论文包含英文摘要和中文摘要，便于进行机器评估。因此，本研究以学术论文摘要作为研究语料，以深入分析常用机器翻译工具的特点。为保证研究语料的权威性及代表性，笔者选取了五部外国语言文学类 CSSCI 期刊，分别为《外国语》《外语教学与研究》《中国外语》《现代外语》及《外语界》，并从每部期刊中随机抽取近三年内发表的六篇论文(共计 30 篇)，收集其中文摘要及作者提供的英文摘要(见附录)。

3.2. 研究工具

笔者选取了有道翻译、DeepL、百度文心一言和 ChatGPT-3.5 四种机器翻译工具对上述 30 篇中文摘要进行英译，获取时间为 2024 年 2 月。在使用大语言模型百度文心一言 2.5.3 版本和 ChatGPT-3.5 时，均提前给定提示语：所翻译的文本为学术论文摘要，请注意用词和句式。翻译完成后使用试译宝译文在线评测工具(https://www.shiyibao.com/tools/MTPEtest)逐个计算论文摘要机器翻译的 BLEU、TER 和 METEOR 评测指标得分，并汇总得出综合得分。

3.3. 数据分析

本研究运用 SPSS 21.0 对收集的数据开展如下分析：

首先，整合不同评估指标得分，计算各译文的综合得分，并对所有机器译文的综合得分进行正态分布检验。若数据符合正态分布，则使用单因素方差分析；否则，使用 Kruskal-Wallis 检验。

鉴于不同译文评估手段的侧重点不同，本研究还将单独对各个指标的分数进行分析，检验流程同上。

3.4. 研究问题

通过对比分析，本研究主要试图回答以下问题：

1) 各机器翻译产品在翻译摘要时质量如何，是否存在局限性？

2) 哪些机器翻译产品在翻译摘要时表现更优？

4. 结果与讨论

4.1. 不同机器翻译译本质量比较

笔者首先对所收集到的 BLEU、TER、METEOR 各项指标进行整合，得出各个译本的综合得分，对综合得分进行正态分布检验。本次测量数据样本量为 120，大于 50，故使用 Kolmogorov-Smirnov 检验，分析结果显示，四组不同机器翻译译文的 p 值分别为 0.200，0.072，0.200 和 0.193，均大于 0.05，四组综合得分符合正态分布，相关数据如表 1 所示。因此，比较不同机器翻译工具的译文质量的综合得分需采用单因素方差分析。

Table 1. Normality test results for overall scores of different machine translation outputs
表 1. 不同机器译本综合得分正态性检验结果

		正态性检验					
	机器翻译工具	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		统计量	df	Sig.	统计量	df	Sig.
综合得分	有道翻译	0.097	30	0.200*	0.972	30	0.583
	DeepL	0.153	30	0.072	0.930	30	0.049
	文心一言	0.124	30	0.200*	0.964	30	0.385
	ChatGPT	0.132	30	0.193	0.917	30	0.023

*. 这是真实显著水平的下限。a. Lilliefors 显著水平修正。

不同机器翻译工具的译文质量比较结果如表 2 和表 3 所示。

Table 2. One-way ANOVA results for overall scores of machine translation outputs
表 2. 机器译本综合得分单因素方差分析结果

单变量检验						
因变量：综合得分						
	平方和	df	均方	F	Sig.	偏 Eta 方
对比	135.825	3	45.275	0.874	0.457	0.022
误差	6007.100	116	51.785			

F 检验机器翻译工具的效应。该检验基于估算边际均值间的线性独立成对比较。

Table 3. Comparison results of overall scores across different machine translation outputs
表 3. 不同机器译本综合得分比较结果

多个比较						
因变量：综合得分						
Bonferroni						
(I) 机器翻译工具	(J) 机器翻译工具	均值差值(I-J)	标准误差	Sig.	95%置信区间	
					下限	上限
有道翻译	DeepL	1.9000	1.85805	1.000	−3.0875	6.8875
	文心一言	−1.0000	1.85805	1.000	−5.9875	3.9875
	ChatGPT	−0.2000	1.85805	1.000	−5.1875	4.7875
DeepL	有道翻译	−1.9000	1.85805	1.000	−6.8875	3.0875
	文心一言	−2.9000	1.85805	0.728	−7.8875	2.0875
	ChatGPT	−2.1000	1.85805	1.000	−7.0875	2.8875
文心一言	有道翻译	1.0000	1.85805	1.000	−3.9875	5.9875
	DeepL	2.9000	1.85805	0.728	−2.0875	7.8875
	ChatGPT	0.8000	1.85805	1.000	−4.1875	5.7875
ChatGPT	有道翻译	0.2000	1.85805	1.000	−4.7875	5.1875
	DeepL	2.1000	1.85805	1.000	−2.8875	7.0875
	文心一言	−0.8000	1.85805	1.000	−5.7875	4.1875

基于观测到的均值。误差项为均值方(错误) = 51.785。

方差分析表明，四个机器翻译工具所翻译的外国语言文学类核心期刊论文摘要英文译文质量不存在显著差异($F = 0.874$, $p = 0.457 > 0.05$)；Bonferroni 事后检验结果表明四个机器翻译工具翻译的译文不存在显著差异，但其中 DeepL 翻译的文本和百度文心一言所翻译的文本存在一定程度的差异趋势。

由表 4 可知，百度文心一言所翻译的摘要英文译本的评分平均值最高，其次是 ChatGPT 与有道翻译，DeepL 的得分均值最低。由此可见，在机器评估体系中，文心一言所翻译的学术论文摘要质量最好，ChatGPT 次之，有道翻译位列第三，DeepL 表现最差。ChatGPT 翻译的摘要分值标准差在四种机器翻译中最小，可见 ChatGPT 的翻译较为稳定。其余三个机器翻译的标准差

数值较为接近，而其中 DeepL 所翻译的英文摘要分值标准差最大，提示其翻译质量波动性较高，对同一学科内不同论文摘要的翻译表现不均。

Table 4. Descriptive statistics for overall scores of different machine translation outputs
表 4. 不同机器译本综合得分描述性统计量

描述性统计量			
因变量：综合得分			
机器翻译工具	均值	标准偏差	N
有道翻译	45.4000	7.19003	30
DeepL	43.5000	7.37072	30
文心一言	46.4000	7.32309	30
ChatGPT	45.6000	6.89127	30
总计	45.2250	7.18479	120

4.2. 不同机器翻译译本 BLEU 指标比较

笔者将不同机器翻译文本的 BLEU 指标进行正态分布检验。本次测量数据样本量为 120，大于 50，故使用 Kolmogorov-Smirnov 检验，分析结果显示，四组的 p 值分别为 0.200，0.200，0.200 和 0.020，有三组数值大于 0.05，有一组数值小于 0.05，相关数据如表 5 所示。由于 SPSS 对数据比较敏感，而此处每组数据量只有 30 个，数据量不大，因此观察数据是否满足正态分布还需结合图形结果，如图 1 的 Q-Q 图所示，ChatGPT 的 BLEU 分值基本符合正态分布。四组 BLEU 数值均符合正态分布。因此，比较不同机器翻译工具的译文的 BLEU 指标需采用单因素方差分析。

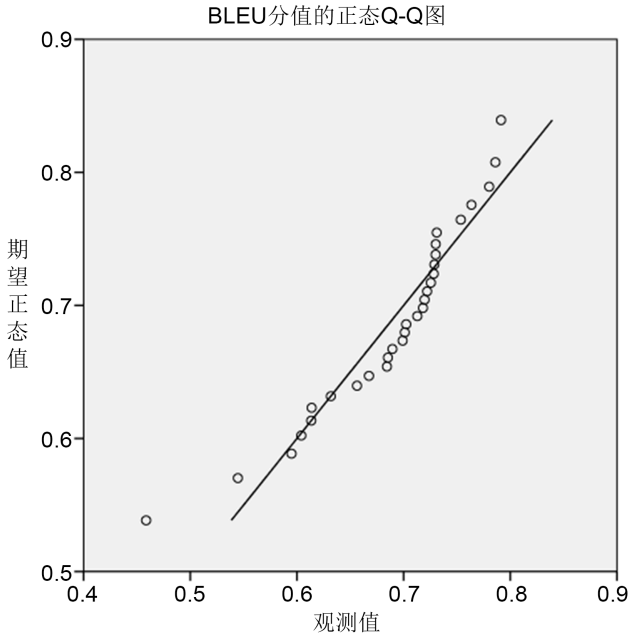

Figure 1. Standard Q-Q plot of BLEU scores for ChatGPT outputs
图 1. ChatGPT 的 BLEU 分值标准 Q-Q 图

Table 5. Normality test results for BLEU scores of different machine translation outputs
表 5. 不同机器译本 BLEU 指标正态性检验结果

正态性检验							
	机器翻译工具	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		统计量	df	Sig.	统计量	df	Sig.
BLEU 分值	有道翻译	0.089	30	0.200*	0.954	30	0.222
	DeepL	0.090	30	0.200*	0.977	30	0.732
	文心一言	0.108	30	0.200*	0.953	30	0.207
	ChatGpt	0.175	30	0.020	0.902	30	0.009

*. 这是真实显著水平的下限。a. Lilliefors 显著水平修正。

BLEU 指标的单因素方差分析结果表明，显著性为 0.164，大于 0.05，所以得出结论不同机器翻译的英文摘要 BLEU 指标不存在着统计学差异(见表 6)。

Table 6. One-way ANOVA results for BLEU scores of different machine translation outputs
表 6. 不同机器译本 BLEU 指标单因素方差分析结果

单变量检验						
因变量：BLEU 分值						
	平方和	df	均方	F	Sig.	偏 Eta 方
对比	0.027	3	0.009	1.735	0.164	0.043
误差	0.610	116	0.005			

F 检验机器翻译工具的效应。该检验基于估算边际均值间的线性独立成对比较。

由表 7 可知，四个机器翻译工具中文心一言的均值最高，约为 0.7021，有道翻译和 ChatGPT 的均值十分接近，差值约为 0.005，而 DeepL 的均值与其他三个机器翻译工具的均值相比相差较大，仅为 0.6619。BLEU 指标主要衡量参考译文与被测评译文之间的相似度，此算法会分别统计参考译文与测评译文中各个单词出现的频率并进行比较。由此可见文心一言大语言模型翻译的英文摘要的用词与外国语言文学类期刊论文英文摘要的作者用词习惯更为接近。

Table 7. Descriptive statistics for BLEU scores of different machine translation outputs
表 7. 不同机器译本 BLEU 指标描述性统计量

描述性统计量			
因变量：BLEU 分值			
机器翻译工具	均值	标准偏差	N
有道翻译	0.694263	0.0675970	30
DeepL	0.661927	0.0775110	30
文心一言	0.702090	0.0709058	30
ChatGPT	0.688917	0.0737170	30
总计	0.686799	0.0731933	120

4.3. 不同机器翻译译本 TER 指标比较

笔者将不同机器翻译文本的 TER 指标进行正态分布检验。本次测量数据样本量为 120，大于 50，故使用 Kolmogorov-Smirnov 检验，分析结果显示，四组的 p 值分别为 0.200，0.200，0.005 和 0.016，有两组数值大于 0.05，有两组数值小于 0.05，相关数据如表 5 所示。由于 SPSS 对数据比较敏感，而此处每组数据量只有 30 个，数据量不大，因此观察数据是否满足正态分布还需结合图形结果，如图 2 和图 3 的 Q-Q 图所示，文心一言和 ChatGPT 的 TER 分值基本符合正态分布。四组 TER 数值符合正态分布。因此，比较不同机器翻译工具的译文的 TER 指标需采用单因素方差分析(见表 8)。

Table 8. Normality test results for TER scores of different machine translation outputs
表 8. 不同机器译本 TER 指标正态性检验结果

		正态性检验					
	机器翻译工具	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		统计量	df	Sig.	统计量	df	Sig.
TER 分值	有道翻译	0.129	30	0.200*	0.943	30	0.109
	DeepL	0.089	30	0.200*	0.959	30	0.294
	文心一言	0.197	30	0.005	0.912	30	0.017
	ChatGPT	0.178	30	0.016	0.898	30	0.007

*. 这是真实显著水平的下限。a. Lilliefors 显著水平修正。

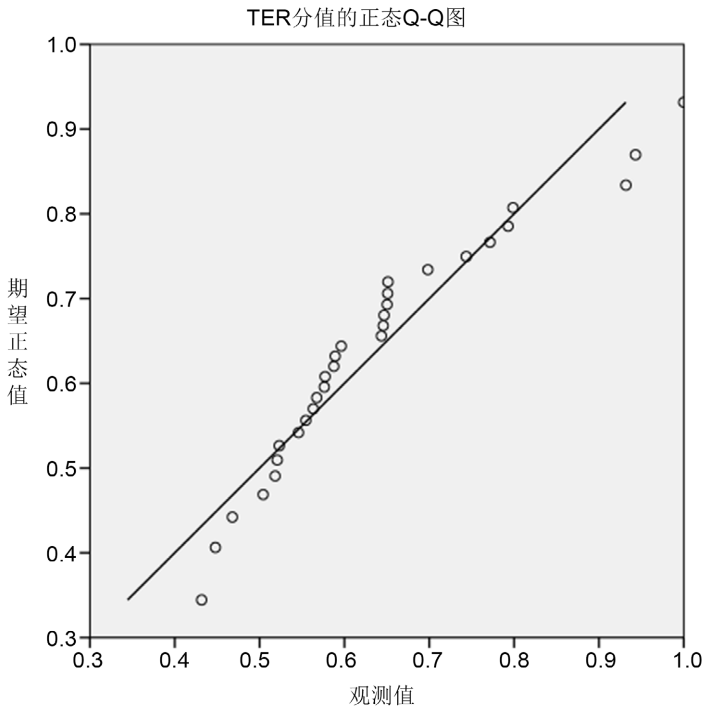

Figure 2. Standard Q-Q plot of TER scores for ERNIE bot outputs
图 2. 文心一言的 TER 分值标准 Q-Q 图

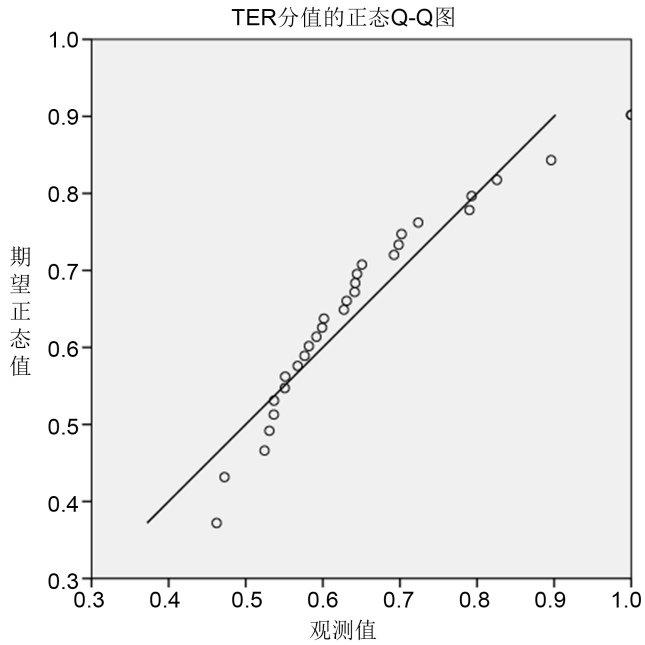

Figure 3. Standard Q-Q plot of TER scores for ChatGPT outputs
图 3. ChatGPT 的 TER 分值标准 Q-Q 图

如表 9 所示，TER 指标的单因素方差分析结果表明，显著性为 0.169，大于 0.05，所以得出结论不同机器翻译的英文摘要 TER 指标不存在着统计学差异。

Table 9. One-way ANOVA results for TER scores of different machine translation outputs
表 9. 不同机器译本 TER 指标单因素方差分析结果

单变量检验						
因变量：TER 分值						
	平方和	df	均方	F	Sig.	偏 Eta 方
对比	0.122	3	0.041	1.709	0.169	0.042
误差	2.766	116	0.024			

F 检验机器翻译工具的效应。该检验基于估算边际均值间的线性独立成对比较。

由表 10 可知，四个机器翻译工具中 DeepL 的均值最高，约为 0.7230，有道翻译和 ChatGPT 的均值较为接近，差值约为 0.0127，文心一言的均值最低，约为 0.6380。TER 指标评估的是机器译文修改为参考译文的后编辑次数，此算法可以处理长句子和重复单词的情况。TER 指标的数值越低，译文质量越高，结合表中数据可见，文心一言的表现最佳。但是该指标的局限是不能处理同义词和短语重组的情况。在所有译本中，有 9 篇译文的 TER 数值为 1，其中摘要 1 的四个机器翻译文本数值均为 1。该摘要的中英行文对比语序调整空间大，但机器翻译追求字面对等，未能灵活调整语序以适应英文表达习惯。这表明，尽管机器翻译技术持续演进，大语言模型亦已兴起，其在处理语篇衔接与连贯方面的灵活性仍有待提升。

Table 10. Descriptive statistics for TER scores of different machine translation outputs
表 10. 不同机器译本 TER 指标描述性统计量

描述性统计量			
因变量：TER 分值			
机器翻译工具	均值	标准偏差	N
有道翻译	0.667383	0.1550652	30
DeepL	0.723023	0.1773231	30
文心一言	0.637997	0.1439074	30
ChatGPT	0.654667	0.1385477	30
总计	0.670768	0.1558029	120

4.4. 不同机器翻译译本 METEOR 指标比较

笔者将不同机器翻译文本的 METEOR 指标进行正态分布检验。本次测量数据样本量为 120，大于 50，故使用 Kolmogorov-Smirnov 检验，分析结果显示，四组的 p 值分别为 0.200，0.116，0.200 和 0.192，四组数值均大于 0.05，符合正态分布，相关数据如表 11 所示。因此，比较不同机器翻译工具的译文的 METEOR 指标需采用单因素方差分析。

Table 11. Normality test results for METEOR scores of different machine translation outputs
表 11. 不同机器译本 METEOR 指标正态性检验结果

正态性检验							
	机器翻译工具	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		统计量	df	Sig.	统计量	df	Sig.
METEOR 分值	有道翻译	0.107	30	0.200*	0.976	30	0.699
	DeepL	0.144	30	0.116	0.924	30	0.034
	文心一言	0.091	30	0.200*	0.983	30	0.900
	ChatGpt	0.132	30	0.192	0.923	30	0.032

*. 这是真实显著水平的下限。a. Lilliefors 显著水平修正。

如表 12 所示，METEOR 指标的单因素方差分析结果表明，显著性为 0.913，大于 0.05，故认为不同机器翻译的英文摘要 METEOR 指标无统计学差异。

Table 12. One-way ANOVA results for METEOR scores of different machine translation outputs
表 12. 不同机器译本 METEOR 指标单因素方差分析结果

单变量检验						
因变量：METEOR 分值						
	平方和	df	均方	F	Sig.	偏 Eta 方
对比	0.001	3	0.000	0.175	0.913	0.005
误差	0.287	116	0.002			

F 检验机器翻译工具的效应。该检验基于估算边际均值间的线性独立成对比较。

由表 13 可知，四个机器翻译工具的 METEOR 分值较为接近，有道翻译和 DeepL 及文心一言三者的数值十分接近，ChatGPT 分值与其他三个相比稍低，低于平均值约 0.0055，差异较小。METEOR 考虑了句子流畅性，同义词对语义的影响，和 BLEU 相比更能反映人类对文本质量的评价，因为它能更灵活地处理单词匹配和词序问题。在本研究中，所有译本的 METEOR 值都偏低，这一指标的数值区间为 0~1，但是在所有译本的数值基本都在 0.3 上下，在四个机器翻译工具所翻译的 120 个译文文本中，仅有四个译文文本的 METEOR 数值高于 0.4，分别为 DeepL 和有道翻译所翻译的摘要 3，DeepL 翻译的摘要 9 和文心一言翻译的摘要 22，反映出机器翻译在达到人类级别流畅度方面仍面临挑战。

Table 13. Descriptive statistics for METEOR scores of different machine translation outputs
表 13. 不同机器译本 METEOR 指标描述性统计量

描述性统计量			
因变量：METEOR 分值			
机器翻译工具	均值	标准偏差	N
有道翻译	0.311020	0.0490087	30
DeepL	0.312173	0.0527593	30
文心一言	0.313707	0.0426479	30
ChatGPT	0.305027	0.0538228	30
总计	0.310482	0.0492321	120

METEOR 指标以词语为单位测评，对词语形式的相似性最为敏感。然而，专业论文中常出现缩略词，例如将 second language 简写成为 L2。而 METEOR 指标无法有效识别这类词语，这也是此次研究中论文英文摘要译本 METEOR 数值偏低的原因之一。例如，第九篇摘要涉及二语习得，参考译文中将大部分的二语写作译为 L2 writing，仅在首次提及时译为 second language writing，而在四个机器译本中只有有道翻译在翻译这一名词时使用了缩略词，且处理方式与参考译文相同，在首次完整翻译后便只使用 L2 这一表达。在第 11 篇摘要中同样也涉及了二语，只有文心一言翻译出了 L2，且主动进行了说明补充。摘要中文为“……首次考察心境和词汇类型对二语情绪词加工的影响……”，文心一言对应译文为“……this study examined the effects of mood and vocabulary type on the processing of second language (L2) emotional words……”。第 19、第 21、第 22、第 24 和第 30 篇摘要中同样也与二语有关，文心一言和有道翻译的对应译本中都有四篇处理成为 L2，而其他两个机器翻译译文均没有出现，可见文心一言及有道翻译在翻译专业文本时对缩略词的翻译表现更好。第二篇摘要中提到了 Readings in Chinese Literary Thought 的作者宇文所安，他的英文名为 Stephen Owen，而四个机器翻译均将其名译为 Yuwen Suoan，出现了偏差。值得注意的是，文心一言与 ChatGPT 作为交互式大语言模型，理论上在专有名词翻译上应更具优势，但实际表现未尽如人意，表明当前大语言模型在专业领域的知识精确性仍有提升空间。

5. 结语

本研究围绕四种主流机器翻译系统在外国语言文学类论文摘要英译任务中的表现展开评估。研究结果表明，有道翻译，DeepL，文心一言及 ChatGPT-3.5 四个机器翻译在翻译外国语言文学类论文摘要时，其综合质量与各自动评估指标上均未呈现显著差异。具体而言，文心一言在多项指标中表现较优，ChatGPT-3.5 的译文质量最为稳定，有道翻译与 DeepL 则各有长短。然而，所有系统在语序灵活性、专业术语处理及文化专有项转换等方面，均与人工翻译存在明显差距，尤其在处理文学理论概念、作者专名及学科特定缩略语时，暴露出语义深层理解与领域适应性的不足。

展望未来，机器翻译技术的发展或可呈现以下趋势：首先，上下文感知与文体适应能力将不断增强。随着大语言模型技术的深度融合，机器翻译系统有望更精准地识别学术文本的语体特征与学科规范，从而生成在风格上更贴近目标语学术惯例的译文。其次，人机协同的译后编辑模式将更为普及。将高质量机器翻译与专业译者的审校能力相结合，形成高效闭环，是提升学术翻译质量与效率的可行路径。再者，多模态与跨语言知识融合将成为重要方向。通过引入图像、语音等辅助信息，系统可加强对文化负载词、专有名词的理解，提升翻译的准确性与地道性。最后，领域自适应能力将通过持续学习与微调机制得到强化，使机器翻译系统能更好地掌握特定学科的术语体系与表达惯例。

诚然，本研究还存在一定的局限性，如语料规模有限、评估完全依赖自动指标等。未来研究可进一步扩大样本数量，并引入人工质量评估，结合更多质性分析，更为深入地揭示机器翻译在学术文本处理中的具体问题与发展路径，为推动机器翻译技术在学术交流中的有效应用提供更全面的依据。

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] 冯志伟. 机器翻译——从梦想到现实[J]. 中国翻译, 1999(4): 38-41.

[2] Hovy, E., King, M. and Popescu-Belis, A. (2002) Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, **17**, 43-75. https://doi.org/10.1023/a:1025510524115

[3] Haque, R., Hasanuzzaman, M. and Way, A. (2020) Analysing Terminology Translation Errors in Statistical and Neural Machine Translation. *Machine Translation*, **34**, 149-195. https://doi.org/10.1007/s10590-020-09251-z

[4] Hudelson, P. and Chappuis, F. (2024) Using Voice-to-Voice Machine Translation to Overcome Language Barriers in Clinical Communication: An Exploratory Study. *Journal of General Internal Medicine*, **39**, 1095-1102. https://doi.org/10.1007/s11606-024-08641-w

[5] Koplenig, A. and Wolfer, S. (2023) Languages with More Speakers Tend to Be Harder to (Machine-)Learn. *Scientific Reports*, **13**, Article No. 18521. https://doi.org/10.1038/s41598-023-45373-z

[6] 黎斌, 唐跃勤. There be 句型在机器翻译软件中的对比研究[J]. 西南交通大学学

报(社会科学版), 2005(2): 84-87.

[7] 黄海英, 冯剑军. 英汉专业翻译软件翻译质量的人工测评[J]. 中国科技翻译, 2008(1): 28-32.

[8] 潘幼博. 人工翻译是否将被机器替代? [J]. 中国科技翻译, 1990(1): 19-23.

[9] 魏长宏, 张春柏. 机器翻译的译后编辑[J]. 中国科技翻译, 2007(3): 22-24+9.

[10] 崔启亮, 李闻. 译后编辑错误类型研究——基于科技文本英汉机器翻译[J]. 中国科技翻译, 2015, 28(4): 19-22.

[11] 冯全功, 刘明. 译后编辑能力三维模型构建[J]. 外语界, 2018(3): 55-61.

[12] 冯志伟, 张灯柯. 机器翻译与人工翻译相辅相成[J]. 外国语(上海外国语大学学报), 2022(6): 77-87.

[13] 陈磊, 李泽世, 王鹏晓, 瞿灼芮, 曹晨宇, 刘红江. 学术论文标题翻译: 常用机器翻译质量评估指标的局限性与多维量化评估标准的构建[J]. 科技传播, 2023, 15(16): 43-46+50.

Appendix (Abstract and Keywords in Chinese)

学术论文摘要机器翻译质量评估研究：以外国语言文学类论文为例

摘要：机器翻译作为人工智能领域的重要应用之一，在跨语言交流与信息传播中发挥着日益关键的作用。近年来，随着神经机器翻译与大语言模型的快速发展，机器翻译的质量与适用场景得到了显著拓展。然而，机器翻译系统的质量直接影响着信息的准确传达和理解。本研究旨在探讨不同机器翻译系统在翻译外国语言文学类论文摘要的翻译质量。本文选取有道翻译，DeepL，文心一言和 ChatGPT-3.5 作为研究对象，对上述四种机器翻译工具所翻译的学术论文摘要进行评估与分析，探究其在学术文本翻译中的具体表现与局限。

关键词：机器翻译，译文质量，论文摘要，评估